D828–D833 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 27 November 2012*
doi:10.1093/nar/gks1231

PrePPI: a structure-informed database of protein–protein interactions

Qiangfeng Cliff Zhang[1], Donald Petrey[1], José Ignacio Garzón[1], Lei Deng[1,2] and Barry Honig[1,*]

[1]Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, Center for Computational Biology and Bioinformatics, Columbia Initiative in Systems Biology, Columbia University, New York, NY 10032, USA and [2]School of Software, Central South University, Changsha 410083, China

Received September 19, 2012; Revised October 29, 2012; Accepted October 31, 2012

ABSTRACT

PrePPI (http://bhapp.c2b2.columbia.edu/PrePPI) is a database that combines predicted and experimentally determined protein–protein interactions (PPIs) using a Bayesian framework. Predicted interactions are assigned probabilities of being correct, which are derived from calculated likelihood ratios (LRs) by combining structural, functional, evolutionary and expression information, with the most important contribution coming from structure. Experimentally determined interactions are compiled from a set of public databases that manually collect PPIs from the literature and are also assigned LRs. A final probability is then assigned to every interaction by combining the LRs for both predicted and experimentally determined interactions. The current version of PrePPI contains ~2 million PPIs that have a probability more than ~0.1 of which ~60 000 PPIs for yeast and ~370 000 PPIs for human are considered high confidence (probability > 0.5). The PrePPI database constitutes an integrated resource that enables users to examine aggregate information on PPIs, including both known and potentially novel interactions, and that provides structural models for many of the PPIs.

INTRODUCTION

Knowledge of protein–protein interactions (PPIs) is essential to understanding cellular regulatory processes. Much effort involving a multitude of methods has been devoted to the determination of direct physical interactions between proteins (1,2). Although most detection methods can only be used for small-scale studies, a few techniques, such as the yeast two-hybrid assays and affinity purification, can be scaled up to determine PPIs in a high-throughput manner (3,4). These high-throughput techniques have been applied to genome-wide studies of PPIs for a number of model organisms, including yeast (5–12), fly (13), worm (14), bacteria (15,16), human (17–19) and, more recently, *Arabidopsis* (20).

A number of databases have been created to systematically collect and store information on experimentally determined PPIs, including the Munich Information Center for Protein Sequence (MIPS) protein interaction database (21), the database of interacting proteins [DIP, (22)], the protein interaction database [IntAct, (23)], the molecular interaction database [MINT, (24)], the Human Protein Reference Database [HPRD, (25)] and the Biological General Repository for Interaction Datasets [BioGRID, (26)]. To date, hundreds of thousands of PPIs have been stored in these databases that cover hundreds of different organisms and contain interactions determined by tens of different methods (27,28).

Although these databases are crucially valuable resources, they inevitably contain some number of false interactions (false positives) and are largely incomplete in that many interactions are still not annotated (false negatives) (29–31). Although false negatives mainly result from the inherent limitations of different detection methods and incomplete screening of the vast possible interaction space, false positives in these databases can result from errors or ambiguities in experiments (32). In particular, data sets generated from high-throughput methods are estimated to have a much higher error rate than traditional small-scale studies (33). In addition to experimental errors, false-negative and false-positive interactions also result from curation errors. For example, a study of discrepancies between different databases showed that, even for the same set of publications, two databases on average only fully agree on 42% of the interactions and 62% of the proteins (34). The differences were attributed to divergent assignments of organism or splice isoforms, and alternative representations of multiprotein complexes, etc.

Parallel to experimental studies and literature curations, computational predictions have also been used to infer

*To whom correspondence should be addressed. Tel: +1 212 851 4651; Fax: +1 212 851 4650, Email: bh6@columbia.edu

new interactions from indirect clues. Information such as sequence and structural homology, domain–domain interaction profile, genomic context, gene fusion, phylogenetic profile/tree similarity, gene co-expression, function similarity and network topology has been effectively exploited to evaluate the reliabilities of experimentally determined interactions (35,36), and to predict PPIs on a large scale (37–41). Usually, every indirect clue by itself is only a weak PPI predictor, but predictions can be improved by integrating different sources of evidence using a variety of machine learning methods. There have been a number of online databases that store PPIs predicted from these integrative methods, such as STRING (42), Predictome (43), OPHID (44) and its replacement I2D, IntNetDB (45) and PIPs (46). These databases have their own limitations, and it should be noted that, owing to the nature of many prediction methods, many of the predicted interactions are often more indicative of protein functional associations than of direct physical interactions.

Recently, we described a PPI prediction method (PrePPI) that is largely based on 3D protein structural information (47). We showed that, with the exploitation of homology models and remote geometric relationships, structural information can be used to accurately predict PPIs on a genome-wide scale. The further integration of structural with other functional clues yields prediction performance comparable with high-throughput experiments. Experimental tests of a number of predictions demonstrate the ability of the structure-based algorithm to identify novel unsuspected PPIs of significant biological interest.

Given the inconsistent levels of reliability and lack of complete overlap between different PPI databases, a resource that integrates different sources of information and that reports an appropriate measure of reliability should be extremely valuable. In this article, we describe the PrePPI database that contains interactions predicted from our structure-based integrative method, and also includes interactions compiled from a set of public databases that manually curate experimentally determined PPIs from the literature. A probability for each interaction is calculated using a Bayesian framework as described later in the text.

THE PrePPI DATABASE: DATA SOURCES

Predicted interactions

Predicted interactions in the PrePPI database are generated by our structure-based integrative PPI prediction method that combines structural modeling with other genomic, evolutionary and functional clues (47). Briefly, for a pair of proteins of interest, we first search for representative structures of the query proteins in the PDB and homology model databases, and then use these to search for structural neighbors of each protein. A protein–protein complex found in the Protein Quaternary Structure database or Protein Data Bank is used as a 'template' for the interaction whenever it contains a pair of interacting chains that are structural neighbors of the respective query proteins. We then construct a model by superposing the individual subunits on their corresponding structural neighbors in the template complex and calculate a likelihood ratio (LR) for each model to represent a true interaction using a Bayesian network trained on a positive and a negative interaction reference set. We finally combine the structure-derived LR with non-structural evidence associated with the query proteins using a naïve Bayesian classifier.

Our analyses show that the performance of the prediction method is comparable with high-throughput studies, and that this is primarily due to the large-scale use of structural information made possible by the use of homology models and looking broadly across protein structure space for structure/function relationships. To put this in perspective, using structure alone we build structural models for ~2.4 million and 36 million yeast and human interactions, respectively.

Experimentally determined interactions

We collected PPIs from six publicly available databases (MIPS, DIP, IntAct, MINT, HPRD and BioGRID) and obtained 117 803 interactions for yeast and 82 060 interactions for human. We mapped protein identifiers from different databases to UniProt accession numbers and used pairs of accession numbers as the unique identifiers of all PPIs. Different databases contain different numbers of false-positive and false-negative interactions that are due to both experimental and curation errors. We have used Bayesian statistics to calculate an LR for database interactions as follows. We used a positive reference set that contains 11 851 yeast interactions and 7409 human interactions that have more than one supporting publication, and a negative reference set constructed by pairing proteins located in different cellular compartments (47). We assigned each of these interactions to one of seven categories and calculated an LR for each category. The first category contains interactions that are present in multiple databases, and the other six contain interactions present in exclusively one of the databases listed earlier in the text. In this way, we obtain an objective evaluation that accounts for both experimental and curation quality.

Combining the LRs for predicted and experimentally determined interactions

An advantage of using a Bayesian framework to calculate an LR for each database is that we can easily combine experimentally determined interactions with computationally predicted interactions. Because the two are weakly correlated, we use a naïve Bayesian classifier to combine them by simply multiplying the two LR scores to obtain a combined LR score for each interaction.

In the PrePPI database, we have scaled the combined LR to a probability using the following equation:

$$probability = \frac{LR}{LR + LR_{cutoff}} \quad (1)$$

We use an LR_{cutoff} of 600, which roughly corresponds to a false-positive rate of 0.001, based on the assumption that the probability that an interaction of LR 600 is true is 0.5 (47,48).

The PrePPI database now contains ~2 million PPIs with a probability >0.1. Of these, 61 720 PPIs for yeast and 372 545 PPIs for human have a probability >0.5.

THE PrePPI DATABASE: WEB INTERFACE

The PrePPI database can be queried through the UniProt accession number (e.g. P03989), gene name (e.g. *PRNP*) or protein name (e.g. Histone H2A) of a gene or protein. The server will return a description of the query protein, the number of proteins it interacts with and a table with detailed information about each interaction (Figure 1). Each row of the table lists proteins predicted to interact with the query, the sources of information used in the prediction, different LRs and the final combined probability, as well as whether the interaction has been documented in databases or in the literature.

The sources of information used in the prediction are represented by their 'prediction codes'. Details on different types of information can be found in the 'Help' page of the web server. The 'Prediction LR' column shows the LR obtained from the Bayesian network that combines the different sources of structural and non-structural evidence for the interaction represented by their prediction codes [see (47) for details on the types of evidence used]. We also calculate a 'database LR' as described earlier and combine this with the prediction LR to get a final LR, which is shown in the table as a probability (Final prob.) determined from Equation 1. If an interaction has been previously documented, we put the corresponding database symbols in the seventh column and the PubMed links to the description of the relevant experiments in the eighth column.

Interactions are ordered according to their final probabilities. By default, we only show the high confidence predictions (final probability >0.5), but predictions with lower probabilities can be viewed by clicking the link at the bottom right. All interactions for the query protein can be downloaded by clicking the link at the bottom left.

A unique feature of the PrePPI database is the availability of structural interaction models for those PPIs predicted from our structural modeling algorithm. Figure 2 shows an example of an interaction model built for the human TGF-β receptor type-1 (P36897) and the complement component C1q receptor (Q9NPY3), using a homology model from Skybase (49) for Q9NPY3 and exploiting the remote structural relationship between these monomer structures and a designed protein that forms a homodimer (50). Users can investigate the interaction model and generate experimentally testable hypotheses for how the two proteins interact. It is important to emphasize that no structural refinement of PrePPI models is carried out, so they may contain physically unrealistic features such as steric clashes. The structure-based LR for the model is shown in the viewer and, together with the reasonableness of the model itself,

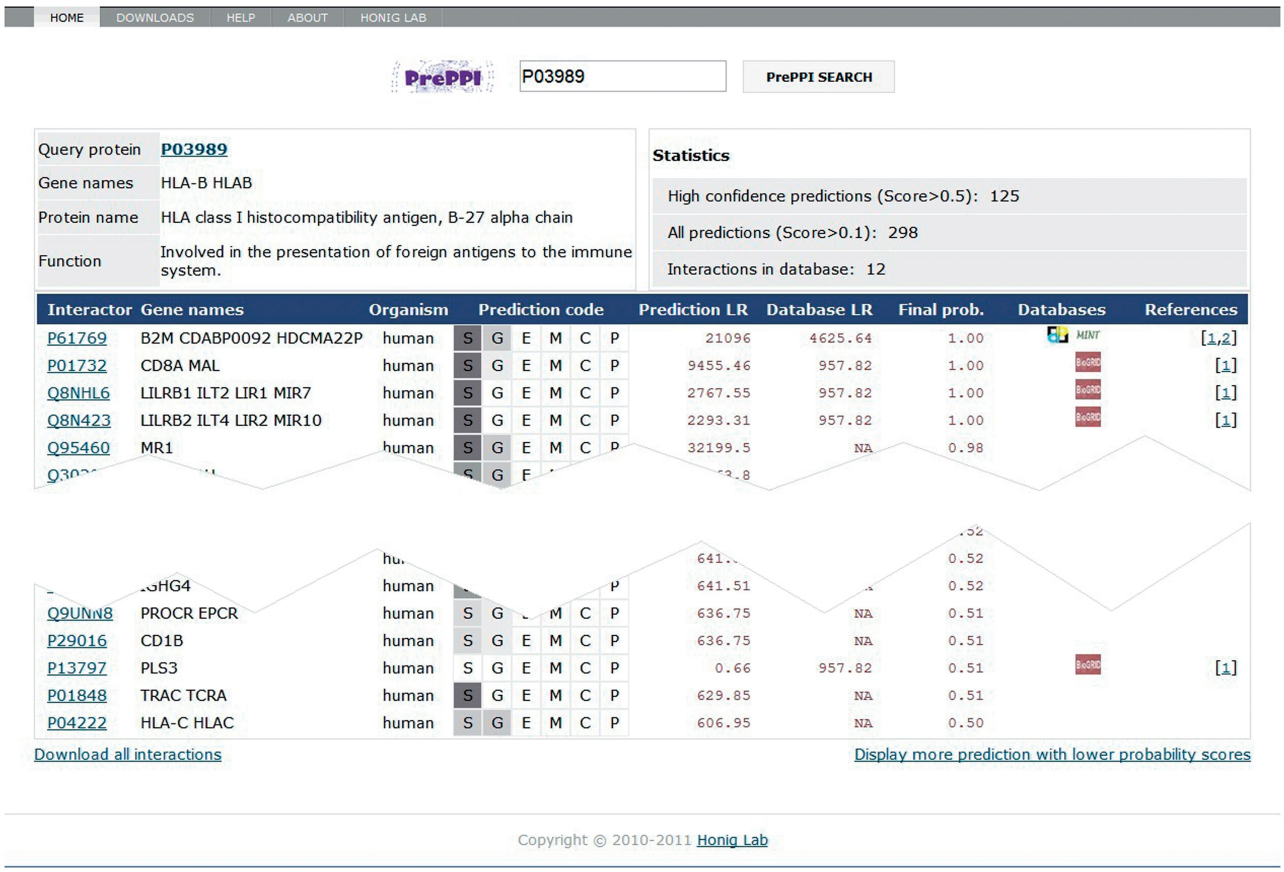

Figure 1. The PrePPI page of predicted protein–protein interactions for query protein P03989.

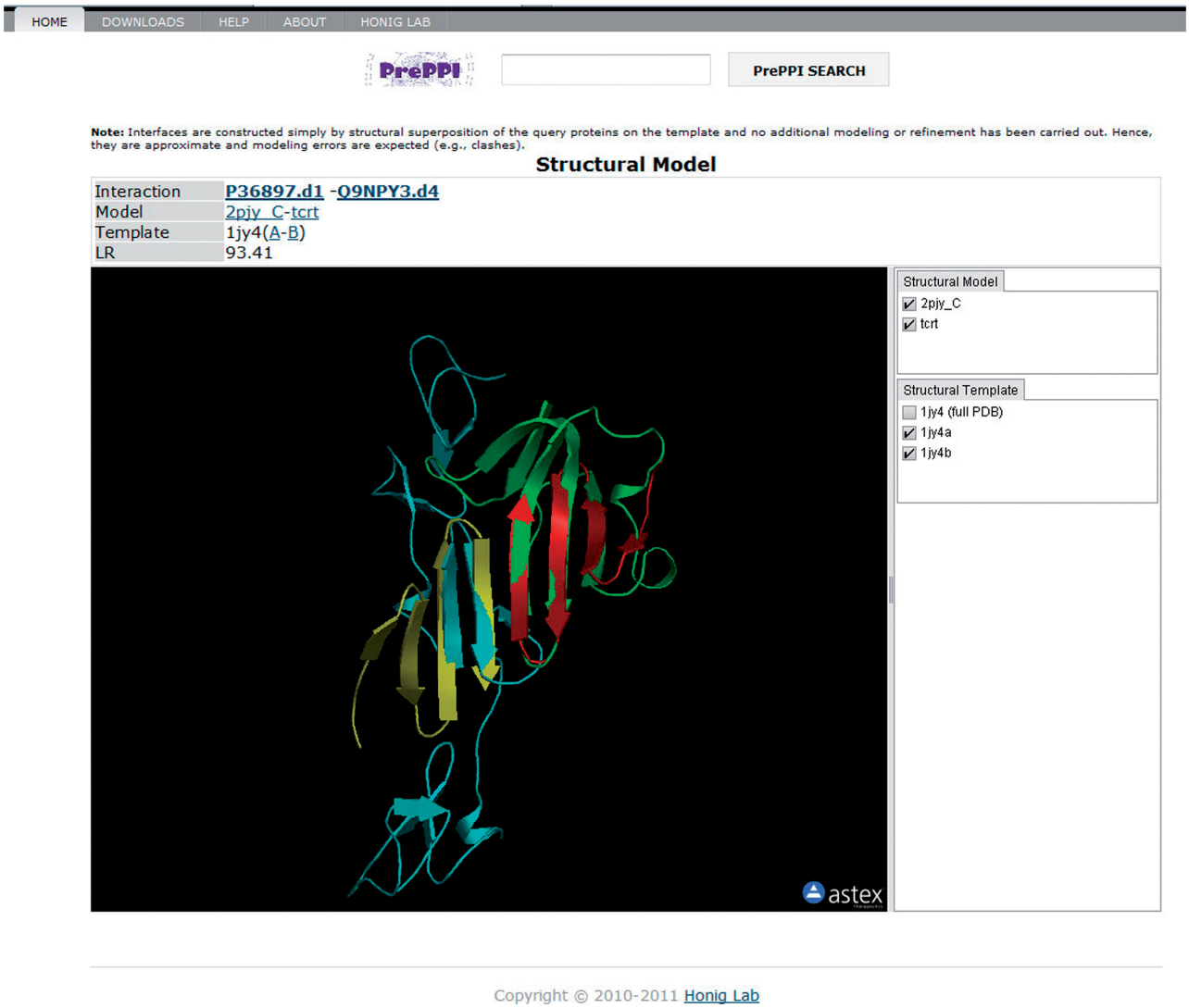

Figure 2. The structural interaction model for TGF-β receptor type I (green, UniProt ID P36897) and complement component C1q receptor (cyan, UniProt ID Q9NPY3) based on the structure of a designed protein (gold and red for A and B chains, respectively, of PDB file 1jy4).

should be considered when evaluating its biological relevance and when deciding whether some form of structural refinement might be of value.

CAVEATS AND PLANS

The goal of PrePPI is to generate testable hypotheses derived in part from structure, but its use should be seen, in our opinion, as an early step in the process of biological discovery. PrePPI is under constant development, but at this stage, it is worth pointing out a number of caveats. First, although we have shown that the structure-based LR can account for specificity in the sense that it can differentiate closely related structural domains that form complexes from those that do not [see Figure S15 in the supplemental material of (47)], the methods used are not perfect and predictions should be considered carefully in the context of any additional data that might be available (for example, the highest scoring predictions may be paralogs that appear in different cellular compartments). As discussed earlier in the text, other problems may arise from the fact that we do not attempt to evaluate the 3D model of a putative complex beyond scoring of the interface (47) so that in many cases the model may appear physically unrealistic. Ideally, it will be possible to address such issues automatically through, for example, the use of orthology databases or refinement of side chains, loops and relative domain orientations. We plan to implement such features in future versions of PrePPI. However, because PrePPI evaluates billions of interaction models (47), structural refinement would have to be carried out in a later filtering step, perhaps motivated by biological interest. At this stage, we have chosen to present all high probability predictions with the expectation that a thoughtful user will be able to recognize obvious false positives using the information available on the server itself, in external databases or in the biological literature.

Finally we note that a high probability PrePPI prediction for an interaction says nothing about the

oligomerization state of the proteins involved. Our goal at this stage is to assign a probability for an interaction between two proteins to occur and provide an initial model of where an interface might be located. Again, our hope is that the interested user will be able to use the information provided in the PrePPI database as a basis for new experimental and computational efforts on a particular system of interest.

CONCLUSION

The PrePPI database differs from other PPI databases based on the following four novel features: (i) PrePPI provides structural information for many more interactions than has previously been possible using structure-enabled approaches and databases (51–53); (ii) the predicted PPIs in PrePPI are obtained by combining structural and non-structural information; (iii) the PrePPI database contains integrative information of PPIs from major PPI databases and provides a Bayesian measure as to the confidence level of these interactions; and (iv) the PrePPI database assigns a single probability for each interaction using a Bayesian framework that combines quantitative results based on computational predictions with evidence contained in publicly available databases. PrePPI now offers a comprehensive source of PPI information for the yeast and human genomes and will soon be expanded to other model organisms.

FUNDING

National Institutes of Health [GM030518, GM094597, CA121852]. Funding of the open access charge: Howard Hughes Medical Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Phizicky,E.M. and Fields,S. (1995) Protein-protein interactions: methods for detection and analysis. *Microbiol. Rev.*, **59**, 94–123.
2. Shoemaker,B.A. and Panchenko,A.R. (2007) Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput. Biol.*, **3**, e42.
3. Parrish,J.R., Gulyas,K.D. and Finley,R.L. Jr (2006) Yeast two-hybrid contributions to interactome mapping. *Curr. Opin. Biotechnol.*, **17**, 387–393.
4. Vasilescu,J. and Figeys,D. (2006) Mapping protein-protein interactions by mass spectrometry. *Curr. Opin. Biotechnol.*, **17**, 394–399.
5. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.
6. Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
7. Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M., Cruciat,C.M. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
8. Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al.* (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.
9. Gavin,A.C., Aloy,P., Grandi,P., Krause,R., Boesche,M., Marzioch,M., Rau,C., Jensen,L.J., Bastuck,S., Dumpelfeld,B. *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.
10. Krogan,N.J., Cagney,G., Yu,H., Zhong,G., Guo,X., Ignatchenko,A., Li,J., Pu,S., Datta,N., Tikuisis,A.P. *et al.* (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, **440**, 637–643.
11. Yu,H., Braun,P., Yildirim,M.A., Lemmens,I., Venkatesan,K., Sahalie,J., Hirozane-Kishikawa,T., Gebreab,F., Li,N., Simonis,N. *et al.* (2008) High-quality binary protein interaction map of the yeast interactome network. *Science*, **322**, 104–110.
12. Tarassov,K., Messier,V., Landry,C.R., Radinovic,S., Serna Molina,M.M., Shames,I., Malitskaya,Y., Vogel,J., Bussey,H. and Michnick,S.W. (2008) An in vivo map of the yeast protein interactome. *Science*, **320**, 1465–1470.
13. Giot,L., Bader,J.S., Brouwer,C., Chaudhuri,A., Kuang,B., Li,Y., Hao,Y.L., Ooi,C.E., Godwin,B., Vitols,E. *et al.* (2003) A protein interaction map of Drosophila melanogaster. *Science*, **302**, 1727–1736.
14. Li,S., Armstrong,C.M., Bertin,N., Ge,H., Milstein,S., Boxem,M., Vidalain,P.O., Han,J.D., Chesneau,A., Hao,T. *et al.* (2004) A map of the interactome network of the metazoan C. elegans. *Science*, **303**, 540–543.
15. Butland,G., Peregrin-Alvarez,J.M., Li,J., Yang,W., Yang,X., Canadien,V., Starostine,A., Richards,D., Beattie,B., Krogan,N. *et al.* (2005) Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature*, **433**, 531–537.
16. Kuhner,S., van Noort,V., Betts,M.J., Leo-Macias,A., Batisse,C., Rode,M., Yamada,T., Maier,T., Bader,S., Beltran-Alvarez,P. *et al.* (2009) Proteome organization in a genome-reduced bacterium. *Science*, **326**, 1235–1240.
17. Rual,J.F., Venkatesan,K., Hao,T., Hirozane-Kishikawa,T., Dricot,A., Li,N., Berriz,G.F., Gibbons,F.D., Dreze,M., Ayivi-Guedehoussou,N. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.
18. Stelzl,U., Worm,U., Lalowski,M., Haenig,C., Brembeck,F.H., Goehler,H., Stroedicke,M., Zenkner,M., Schoenherr,A., Koeppen,S. *et al.* (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.
19. Ewing,R.M., Chu,P., Elisma,F., Li,H., Taylor,P., Climie,S., McBroom-Cerajewski,L., Robinson,M.D., O'Connor,L., Li,M. *et al.* (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.*, **3**, 89.
20. Arabidopsis Interactome Mapping Consortium. (2011) Evidence for network evolution in an Arabidopsis interactome map. *Science*, **333**, 601–607.
21. Mewes,H.W., Albermann,K., Heumann,K., Liebl,S. and Pfeiffer,F. (1997) MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res.*, **25**, 28–30.
22. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
23. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A., Huntley,R. *et al.* (2007) IntAct—open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
24. Chatr-aryamontri,A., Ceol,A., Palazzi,L.M., Nardelli,G., Schneider,M.V., Castagnoli,L. and Cesareni,G. (2007) MINT: the molecular interaction database. *Nucleic Acids Res.*, **35**, D572–D574.
25. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database—2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
26. Stark,C., Breitkreutz,B.J., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.

27. Lehne,B. and Schlitt,T. (2009) Protein-protein interaction databases: keeping up with growing interactomes. *Hum. Genomics*, **3**, 291–297.
28. Tsai,J., Rohl,C., Price,Y., Fischer,T.B., Paczkowsk,M. and Zette,M.F. (2006) Cataloging the relationships between proteins. *Mol. Biotechnol.*, **34**, 69–93.
29. von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.
30. Braun,P., Tasan,M., Dreze,M., Barrios-Rodiles,M., Lemmens,I., Yu,H., Sahalie,J.M., Murray,R.R., Roncari,L., de Smet,A.S. *et al.* (2009) An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods*, **6**, 91–97.
31. Deane,C.M., Salwinski,L., Xenarios,I. and Eisenberg,D. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics*, **1**, 349–356.
32. Sprinzak,E., Sattath,S. and Margalit,H. (2003) How reliable are experimental protein-protein interaction data? *J. Mol. Biol.*, **327**, 919–923.
33. Reguly,T., Breitkreutz,A., Boucher,L., Breitkreutz,B.J., Hon,G.C., Myers,C.L., Parsons,A., Friesen,H., Oughtred,R., Tong,A. *et al.* (2006) Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J. Biol.*, **5**, 11.
34. Turinsky,A.L., Razick,S., Turner,B., Donaldson,I.M. and Wodak,S.J. (2010) Literature curation of protein interactions: measuring agreement across major public databases. *Database*, **2010**, baq026.
35. Deane,C.M., Salwinski,L., Xenarios,I. and Eisenberg,D. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell Proteomics*, **1**, 349–356.
36. Bader,J.S., Chaudhuri,A., Rothberg,J.M. and Chant,J. (2004) Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.*, **22**, 78–85.
37. Shoemaker,B.A. and Panchenko,A.R. (2007) Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput. Biol.*, **3**, e43.
38. Valencia,A. and Pazos,F. (2002) Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol.*, **12**, 368–373.
39. Salwinski,L. and Eisenberg,D. (2003) Computational methods of analysis of protein-protein interactions. *Curr. Opin. Struct. Biol.*, **13**, 377–382.
40. Szilagyi,A., Grimm,V., Arakaki,A.K. and Skolnick,J. (2005) Prediction of physical protein-protein interactions. *Phys. Biol.*, **2**, S1–S16.
41. Musso,G.A., Zhang,Z. and Emili,A. (2007) Experimental and computational procedures for the assessment of protein complexes on a genome-wide scale. *Chem. Rev.*, **107**, 3585–3600.
42. von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.
43. Mellor,J.C., Yanai,I., Clodfelter,K.H., Mintseris,J. and DeLisi,C. (2002) Predictome: a database of putative functional links between proteins. *Nucleic Acids Res.*, **30**, 306–309.
44. Brown,K.R. and Jurisica,I. (2005) Online predicted human interaction database. *Bioinformatics*, **21**, 2076–2082.
45. Xia,K., Dong,D. and Han,J.-D. (2006) IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model. *BMC Bioinformatics*, **7**, 508.
46. McDowall,M.D., Scott,M.S. and Barton,G.J. (2009) PIPs: human protein–protein interaction prediction database. *Nucleic Acids Res.*, **37**, D651–D656.
47. Zhang,Q.C., Petrey,D., Deng,L., Qiang,L., Shi,Y., Thu,C.A., Bisikirska,B., Lefebvre,C., Accili,D., Hunter,T. *et al.* (2012) Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, **490**, 556–560.
48. Jansen,R., Yu,H., Greenbaum,D., Kluger,Y., Krogan,N.J., Chung,S., Emili,A., Snyder,M., Greenblatt,J.F. and Gerstein,M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.
49. Mirkovic,N., Li,Z., Parnassa,A. and Murray,D. (2007) Strategies for high-throughput comparative modeling: applications to leverage analysis in structural genomics and protein family organization. *Proteins*, **66**, 766–777.
50. Venkatraman,J., Nagana Gowda,G.A. and Balaram,P. (2002) Design and construction of an open multistranded β-sheet polypeptide stabilized by a disulfide bridge. *J. Am. Chem. Soc.*, **124**, 4987–4994.
51. Stein,A., Céol,A. and Aloy,P. (2011) 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.*, **39**, D718–D723.
52. Lo,Y.-S., Chen,Y.-C. and Yang,J.-M. (2010) 3D-interologs: an evolution database of physical protein- protein interactions across multiple genomes. *BMC Genomics*, **11**, S7.
53. Davis,F.P. and Sali,A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**, 1901–1907.